arXiv:1103.6038v1 [astro-ph.IM] 30 Mar 2011

Searching for comets on the World Wide Web: The orbit of 17P/Holmes from the behavior of photographers

Dustin Lang[1,2] & David W. Hogg[3,4]

ABSTRACT

We performed an image search on *Yahoo!* for "Comet Holmes" on 2010 April 1. Thousands of images were returned. We astrometrically calibrated—and therefore vetted—the images using the *Astrometry.net* system. The calibrated image pointings form a set of data points to which we can fit a test-particle orbit in the Solar System, marginalizing out image dates and detecting outliers. The approach is Bayesian and the model is, in essence, a model of how comet astrophotographers point their instruments. We find very strong probabilistic constraints on the orbit, although slightly off the JPL ephemeris, probably because of limitations of the astronomer model. Hyper-parameters of the model constrain the reliability of date meta-data and where in the image astrophotographers place the comet; we find that ~ 70 percent of the meta-data are correct and that the comet typically appears in the central $\sim 1/e$ of the image footprint. This project demonstrates that discoveries are possible with data of extreme heterogeneity and unknown provenance; or that the Web is possibly an enormous repository of astronomical information; or that if an object has been given a name and photographed thousands of times by observers who post their images on the Web, we can (re-)discover it and infer its dynamical properties!

Subject headings: celestial mechanics — comets: individual (17P/Holmes) — ephemerides — methods: statistical — surveys — time

[1] Princeton University Observatory, Princeton, NJ, 08544, USA

[2] to whom correspondence should be addressed: dstn@astro.princeton.edu

[3] Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place, New York, NY, 10003, USA

[4] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117, Heidelberg, Germany

1. Introduction

The Web bristles with billions of images: on Web pages, in public photo-sharing sites, on social networks, and in private email and file-sharing conversations. A tiny fraction but *enormous number* of these images are *astronomical* images—images of the night sky in which astronomical sources are visible. This is true even if we exclude from consideration scientific collections such as those of professional observatories and surveys and only count the images of hobbyists, amateurs, and sight-seers. In principle these images, taken together, contain an enormous amount of information about the astronomical sky. Of course they have no scientifically responsible provenance, have never been "calibrated" in any sense of that word, and were (mainly) taken for purposes that are not at all scientific. But having been generated from CCD-like measurements of the intensity field, they cannot help but contain important scientific information. The Web is, therefore, an enormous and virtually unexploited sky survey.

It is difficult to estimate the total number of astronomical images on the Web, and even harder to estimate the total data throughput (*étendu* or equivalent measure of scientific information content). However, by any estimate, it is extremely large. For example, image search results for common astronomical subjects include thousands of astronomical images. The *flickr* photo-sharing site has an astrometry group (administered by the *Astrometry.net* collaboration; more below) with more than 14,000 photos, and its astronomy and astrophotography groups have more than 28,000 and 35,000 respectively. A search for the Orion Nebula on *flickr* returns more than 7000 images, which jointly contain significant information on very faint stars and nebular features. These numbers—derived solely from *flickr* searches—represent only a tiny fraction of the relevant Web content. Of course all these search results contain many non-astronomical images, diagrams, fake data, and duplicates, so use of them for science is non-trivial.

The technical obstacles to making use of Web data are immense: If anything has been learned from our interaction with electronic communication, it is that publisher-supplied or provider-supplied meta-data about Web content are consistently missing, misleading, in error, or obscure. Indeed, when it comes to the astronomical properties of imaging discovered on the Web, most providers don't even know what we want in terms of "meta-data"; we want calibration parameters relating to image date, astrometric coordinate system, photometric sensitivity, and point-spread function, and we want it in machine-readable form. The Virtual Astronomy Multimedia Project (VAMP, Gauthier *et al.* 2008) has defined a format for placing astrometric meta-data in image headers, but the goal of the project is to make "pretty pictures" searchable for education and public outreach purposes, rather than science. They do not consider the problem of producing or verifying calibration meta-data; they assume correct meta-data are provided along with the science images that are used to produce the pretty pictures. Even the Virtual Observatory (`http://ivoa.net`), which concentrates on astronomical meta-data, has no plan for ensuring that meta-data are *correct*, and has no machine-readable form for many quantities of great interest (such as the detailed point-spread-function model); we can't expect the world's amateur astrophotographers to be better organized.

Two important changes are occuring in astronomy that are opening up the possibility that we might exploit data collections as radically confusing as that of the entire Web. The first is that tools are beginning to appear that can perform completely hands-free data analysis tasks. The best example so far is the *Astrometry.net* system, which can take astronomical imaging of completely unknown provenance, and calibrate it astrometrically using the data in the image pixels alone (Lang *et al.* 2010).

The second change is that there has been an enormous increase in the amount and diversity of publicly available professional data—that is, calibrated, trustworthy, science-oriented data in observatory, sky-survey, and individual-investigator collections. These collections are so large and diverse that automated data analysis tools that can trivially interact with extremely heterogeneous data are necessary in many scientific domains. That is, much of the technology required for exploitation of the Web-as-sky-survey is required for *any* mature, data-intensive scientific investigation.

We have been exploring some of these ideas with the *Astrometry.net* project. Not only has the system calibrated thousands of images taken by amateurs and hobbyists, we have interfaced the system with *flickr* (Stumm *et al.*, forthcoming). Users who add an image from their online collection to a group called "astrometry" find the image calibrated automatically by an unmanned bot that downloads the image, calibrates it with *Astrometry.net*, and then posts machine-readable calibration results to the image's page on *flickr*; these have been dubbed "astro-tags". The bot also adds annotations to the image, marking named stars and galaxies from the Messier and NGC/IC catalogs. We make use of the *flickr* Application Programming Interface, something many image and data-sharing sites employ. The success of this suggests that automated maintenance of a heterogeneous crowd-sourced sky survey might be possible in the future.

In this paper, we explore some of the ideas around a Web-as-sky-survey, by performing a scientific investigation of Comet Holmes using Web-discovered, human-viewable (JPEG) images alone. Although we do in principle learn things about Comet Holmes, our main interest is in developing and testing new technologies for observational astrophysics. This project leverages the tendency of humans to point their cameras and telescopes towards interesting things, the ability of *Yahoo!* (or any other search engine) to classify and organize their images, and the ability of *Astrometry.net* to figure out after the fact where they were pointing. What we do is related to other citizen-science projects, like the *GalaxyZoo* (Lintott *et al.* 2011) or the monitoring projects of the *AAVSO* (`http://aavso.org`), except that the participants here are entirely unwitting. We end up showing that science can be done with data taken by citizen observers who know nothing of the scientific goals, and scientists who know nothing of the provenance of any of the observations.

2. Data and calibration

Our data collection began with a search of the World Wide Web. We used the *pYsearch* (Hedstrom 2007) code to access the *Yahoo! Web Search* service[1]. On 2010 April 1, we searched for JPEG-format images using the query phrase "Comet Holmes". This yielded approximately 10,000 total results, but the *Yahoo! Web Search* API allowed only 1000 results to be retrieved per query. In order to broaden the result set, we performed an additional set of searches. For each Web site containing an image in the original set of results, we performed a query that was limited to that Web site. These queries, performed later on 2010 April 16, produced an additional 2741 results (including some duplicates), for a total of 2476 unique results. See Figure 1 for some example images.

Next, we used *wget* to retrieve the images on 2010 April 16. This yielded a total of 2309 valid JPEG images. After removing byte-identical images, 2241 unique images remained. We then ran the *Astrometry.net* code on each image to perform astrometric calibration. 1299 images were recognized as images of the night sky and astrometrically calibrated. These images form the data set we use in our analysis below. Figure 2 shows the footprints of the images on the sky.

Many of the images in this data set are annotated images or diagrams such as finding charts or illustrations of the comet's orbit. Some of these diagrams were recognized by *Astrometry.net* as images of the sky. This can be seen in the co-added image in Figure 2, where there are clearly lines connecting the stars that form the constellation Perseus.

Of the 1299 images in our data set, 422 have timestamps in the image headers ("Exchangeable image file format" or EXIF headers). The distribution of timestamps is shown in Figure 3. On 2007 Oct 24, Comet 17P/Holmes brightened by more than 10 mag (Buzzi *et al.* 2007), generating considerable public interest and making it a very popular and accessible observing target in the amateur astronomy community. The distribution of image

[1] http://developer.yahoo.com/search/web/webSearch.html

timestamps shows a large spike at this time.

We evaluated the accuracy of the image timestamps by asking, for each image, whether the comet would appear within the celestial-coordinate bounds of the image at its stamped time. We find that the majority of the timestamps are consistent, and that inconsistent timestamps are typically late rather than early. See Figure 4.

Figure 3 shows the distribution of angular scales of the images in our data set. The distribution peaks around 3 square degrees. Also shown is the distribution of exposure times reported in the EXIF headers.

3. Orbit inference

We take the approach of generative modeling; that is, we construct a well-defined approximation to the probability of the data given the model. We take the "data" to be the pointing (on the sky) of each astrometrically calibrated image; recall that the goal is to use the *behavior* of astrophotographers (in pointing their cameras) to find the gravitational orbits of objects in the sky. We treat the time at which each image is taken as a hidden parameter, and treat the image field of view (angular size and orientation) to be fixed prior information, established by *Astrometry.net*.

For any image i there is a pointing $\boldsymbol{\alpha}_i$ (two-dimensional position or celestial coordinates on the sky). These are the *data*. The image was taken at time t_i, and has image parameters $\boldsymbol{\Omega}_i$ (camera plate scale, image size, orientation, and reported EXIF timestamp if there is one), taken to be known. In addition, the comet has orbital parameters $\boldsymbol{\omega}$, which can be thought of as semimajor axis, eccentricity, inclination, longitudes, *etc.*, or equivalently a 3-dimensional position $\boldsymbol{x}$ and velocity $\boldsymbol{v}$ at a chosen epoch. We choose the latter for inference simplicity, and use as the epoch JD 2454418.5 (2007 Nov 14). Finally, there are two additional nuisance *hyperparameters* $\boldsymbol{\theta}$ that will appear as we go.

The single-image likelihood marginalized over time t_i is

$$\begin{aligned}
p(\boldsymbol{\alpha}_i|\boldsymbol{\Omega}_i,\boldsymbol{\omega},\boldsymbol{\theta}) &= \int p(\boldsymbol{\alpha}_i|t_i,\boldsymbol{\Omega}_i,\boldsymbol{\omega},\boldsymbol{\theta})\,p(t_i|\boldsymbol{\Omega}_i,\boldsymbol{\theta})\,dt_i \\
p(\boldsymbol{\alpha}_i|t_i,\boldsymbol{\Omega}_i,\boldsymbol{\omega},\boldsymbol{\theta}) &= p_{\mathrm{good}}\,p_{\mathrm{fg}}(\boldsymbol{\alpha}_i|t_i,\boldsymbol{\Omega}_i,\boldsymbol{\omega},\boldsymbol{\theta}) + [1-p_{\mathrm{good}}]\,p_{\mathrm{bg}}(\boldsymbol{\alpha}_i) \\
p_{\mathrm{fg}}(\boldsymbol{\alpha}_i|t_i,\boldsymbol{\Omega}_i,\boldsymbol{\omega},\boldsymbol{\theta}) &= \begin{cases} [\eta\,\Omega_i]^{-1} & \text{comet in } \eta \text{ sub-image} \\ 0 & \text{comet not in } \eta \text{ sub-image} \end{cases} \\
p_{\mathrm{bg}}(\boldsymbol{\alpha}_i) &= [4\pi]^{-1} \quad ,
\end{aligned} \tag{1}$$

where p_{good} is the first hyperparameter in $\boldsymbol{\theta}$ and the probability that the image really *is* a picture intentionally taken of (generated by) the comet, $p_{\rm fg}(\cdot)$ is a "foreground" model, which gives good likelihood when the comet (with orbital parameters $\boldsymbol{\omega}$ at time t_i) is inside the image, $p_{\rm bg}(\cdot)$ is a "background" model (which has no dependence on the comet or time), η is a hyperparameter in $\boldsymbol{\theta}$ (subject to $0 < \eta < 1$) that controls the fractional size of the central region of any image in which astrophotographers place comet subjects, Ω_i is the solid angle covered by image i, the "η sub-image" is the central η of the image, and 4π is the solid angle of the whole sky. In detail, we define the η sub-image to be have the same aspect ratio as the whole image, centered at the same point, but smaller in angular size by $\sqrt{\eta}$ along both dimensions.

The time probability distribution function (PDF) $p(t_i|\boldsymbol{\Omega}_i, \boldsymbol{\theta})$ turns out to be crucial to good inference in this problem, in part because trivial or wrongly uninformative time PDFs lead to highly biased answers, a point to which we will return below. We expect that a large fraction of the image EXIF timestamps (where they exist) are correct, but at the same time we cannot trust them completely. We construct an empirical "cheater" prior $p_{\rm emp}(t)$ based on the empirical histogram of extant EXIF timestamps as follows: We construct a grid of non-overlapping bins in time of width 8 d between $t_{\rm min}$ =2007 July 1 and $t_{\rm max}$ =2008 May 1. We count EXIF timestamps in these bins, and then add 1 to every bin (so no bins have counts of zero). We then normalize so that the integral of $p_{\rm emp}(t)$ is unity. This empirical prior is shown in Figure 5. Given any image i with image parameters $\boldsymbol{\Omega}_i$, the PDF for time t_i is

$$\begin{aligned} p(t_i|\boldsymbol{\Omega}_i, \boldsymbol{\theta}) &= \begin{cases} p_{\rm emp}(t_i) & \text{if no } t_{\rm EXIF} \text{ in } \boldsymbol{\Omega}_i \\ p_{\rm EXIF}\, p(t_i|t_{\rm EXIF}) + [1 - p_{\rm EXIF}]\, p_{\rm emp}(t_i) & \text{if } t_{\rm EXIF} \text{ in } \boldsymbol{\Omega}_i \end{cases} \\ p(t_i|t_{\rm EXIF}) &= \mathrm{uniform}(t_i|t_{\rm EXIF} - [0.5\text{ d}], t_{\rm EXIF} + [0.5\text{ d}]) \quad , \end{aligned} \tag{2}$$

where $p_{\rm EXIF}$ is the third hyperparameter in $\boldsymbol{\theta}$ and the probability that a given EXIF timestamp is reliable, $\mathrm{uniform}(x|A, B)$ is the top-hat or uniform PDF for x between A and B, $t_{\rm EXIF}$ is the reported EXIF timestamp, and we have subtracted and added 0.5 d because the EXIF format contains no time zone information and this is the span of possible time zones! An example is shown in Figure 5.

With this model, a single-image likelihood evaluation (marginalized over time t_i) involves integrating the comet trajectory on a fine time grid, and performing integrals numerically as sums over grid points. For dynamical integration we use Keplerian two-body (Kepler 1609) celestial mechanics code implemented in Python by *Astrometry.net* for both the comet and the Earth–Moon barycenter (EMB); we take the initial conditions of the EMB from JD 2454101.5 (2007 Jan 1). For simplicity we take the EMB to be the observer's location. At the precision of the data, the finite light-travel time in the Solar System is significant; we include it when we consider the observed position of the comet as a function of time. For the numerical integrals, we simply convert observed Solar-system directions to positions on the celestial sphere, and positions on the celestial sphere to image positions (to determine whether particular comet instances are inside particular images) with the *Astrometry.net* world-coordinate system libraries (Lang *et al.* 2010).

The total likelihood is the product of the individual-image marginalized likelihoods, and the posterior PDF for the parameters $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$ is proportional to the total likelihood times a prior. We take this prior to be Gaussian in comet position $\boldsymbol{x}$ with three-dimensional isotropic Gaussian variance of $[1\ \mathrm{AU}]^2$, a beta distribution in squared velocity $v^2 \equiv \boldsymbol{v} \cdot \boldsymbol{v}$ between $v^2 = 0$ and the v^2 that just unbinds the comet, with beta-distribution parameters $\alpha = 1$ and $\beta = 3$. We take the prior to be flat in the range 0 to 1 for the probability hyperparameters p_{good} and p_{EXIF} and (improperly) flat in $\ln(\eta)$ for the fractional hyperparameter η. These 9 parameters (three position components, three velocity components, two probabilities, and one fraction) are the parameters in which we perform our Markov Chain Monte Carlo (MCMC) sampling. We perform the sampling with a Python implementation (Foreman-Mackey & Widrow, forthcoming) of an affine-invariant ensemble sampler (Goodman *et al.* 2010, Hou *et al.*, forthcoming) using an ensemble of 64 walkers and Python multiprocessor support. We initialize the MCMC ensemble in a tiny parameter-space ball around the "true" orbit determined by the Jet Propulsion Lab *Horizons* system (Giorgini *et al.* 1996), with eyeballed-sensible values for p_{good}, p_{EXIF}, and η. This is cheating if we claim discovery; we discuss this further below.

There are several substantial limitations to this model: The prior does not even come close to representing our true prior knowledge about comets, particularly ones that are observed by photographers and posted to the Web. The total likelihood (being a product of individual-image likelihoods) assumes all the data are independent, but in reality some of the images found by the Web search are repeats, duplicates, or derived images from others. Most importantly, we make no attempt to *find the comet in the image*. This is a model of how astronomers point their cameras, not of the visible comet itself.

The results of the inference are shown in Figure 6 as a set of sample trajectories drawn from the Markov chain. These samples are effectively drawn from the posterior PDF marginalized over the hyperparameters. The small dispersion among the samples show that the data—just the pointings of a set of heterogeneous images—are incredibly informative about the comet orbit.

4. Discussion

We have shown that if a Solar System body has been *named* and hundreds of astrophotographers around the world have deliberately *photographed it*, we can recover its dynamical properties by a Web search operation (followed by an obscene amount of computation). All the inference is done on image positions; we never look at the content of the images at all. This effectively makes the model a model of astrophotographers, because the image pointings are a record of where human observers pointed their telescopes and cameras. The six dynamical parameters are parameters of the comet to be sure, but the three hyperparameters are not. The probability p_{good} relates to the purity of image search on the Web (for this relatively Web-unique search term), the probability p_{EXIF} relates to the reliability of astrophotographers' Web-published image meta-data, and the fraction η relates to how astrophotographers frame their images. We find $p_{\mathrm{EXIF}} \sim 0.7$ and $\eta \sim e^{-1}$. We expect that these hyperparameters will vary for different Web search engines, query phrases, and comets: Comets with distinctive names are likely to be better indexed by search engines, faint comets are likely to be photographed by different populations of observers, and comets with long tails are likely to be framed differently in photographs. We like to say that we (re-)discovered the Comet; we didn't really, since we initialized the MCMC at the JPL ephemeris. However, the results do show that there is sufficient information in the imaging to have performed a re-discovery with a good orbital parameter-space search algorithm.

The model is exceedingly crude, and the fact that our results are biased (the samples in Figure 6 are offset from the JPL trajectory) is probably in part related to this crudeness. The centering model is extremely crude; in reality there is a distribution of astrophotographers' behavior that it ought to describe. The time model involves a hard-set empirical prior that is not justified and ought to be simultaneously optimized and marginalized out in the inference (this would be a form of hierarchical inference like in Hogg *et al.* 2010). The time-zone model (flat across all time zones) is also not realistic, since some time zones are much more populated with photographers than others. Along those same lines, there is an enormous amount of external information (weather data and visibility calculations) that could further constrain the possible times and time zones.

Another crudeness is in the assumption of independent and identically distributed draws in the likelihood. This is not true in that some of the Web images we find are crops, edits, or diagrams made from other Web images. That is, each image is not guaranteed to be an independent datum.

In some sense, this project is a citizen-science project, because it does science with data generated by non-scientists. However, it is very different from projects like *SETI@Home* (Korpela *et al.* 2009) because it makes use of participants' intelligence, not just hardware.

It is very different from projects like *GalaxyZoo* (Lintott *et al.* 2011) because it makes use of specialized astronomy knowledge among the participants; you have to be a relatively avid astronomer to usefully contribute. It is very different from the projects of the *AAVSO* (`http://aavso.org`) or *MicroFUN* (Gould 2008) because the observers observed for reasons (for all we know) completely unrelated to our scientific goals. It is different from all of these projects in that the participants contributed unwittingly.

One interesting and ill-understood aspect of a citizen-science project of this type—where the participants don't even know that they are involved—relates to giving proper credit and obtaining proper permissions to use the images. We obtained permission to show the images shown in Figure 1 but we did not even attempt to get any permissions for the majority of the 2241 images we touched in the analysis. One encouraging lesson from this project is that the photographers we *did* contact were very supportive: Not one rejected our request for permissions; typical responses expressed enthusiasm about being involved in a scientific paper; the majority asked to see the manuscript when it appears; some sent updated images or suggestions about which images to use; and a few offered details about the data analysis and processing that was performed. A less encouraging lesson is that Web image search is somewhat disabled, apparently deliberately: The *Yahoo! Web Search* API is being decommissioned; *Google* Image Search substantially limits the size of the results available to the API.

The biggest lesson is that there is enormous information about astronomy available in uncurated non-professional images on the Web. We have only scratched this surface. Think how much better we could have done if we had gone into the images and actually made some attempt at *detecting* the comet! Figure 2 shows that there is far more information inside the images than in just the footprints. Figure 8 shows that there is a similarly informative body of images of Comet C/1996 B2 (Hyakutake). We have also noticed that there are more than 3500 images of the Orion Nebula on *flickr* alone, and thousands more elsewhere on the Web; the joint information in this body of images (about the nebula and about time-domain activity therein) must be staggering. Perhaps this is not surprising given the large amount of glass and detector area owned by avid photographers. We have learned that you can do high-quality quantiative astrophysics with images of unknown provenance on the Web. Is it possible to build from these images a true sky survey? We expect the answer is "yes".

We are very sad that Sam Roweis (Toronto, Google, NYU) was not here to collaborate on this project, after co-creating with us *Astrometry.net*; he influenced every aspect of this project. We obtained ideas, feedback, or code from Michael Blanton (NYU), Jo Bovy (NYU), Daniel Foreman-Mackey (NYU), Jonathan Goodman (NYU), Fengji Hou (NYU), Iain Murray (Edinburgh), and Christopher Stumm (Microsoft). We benefitted from the activity of a very large number of astrophotographers and Web citizens, but in particular we got permissions to show images from John F. Pane, Stephane Zoll, Thorsten Boeckel, Vincent Jacques, Babek Tafreshi, Per Magnus Hedén, Jimmy Westlake, Vicent Peris, Paolo Berardi, Fay Saunders, Dave Kodama, Tyler Allred, Joe Orman, Ivan Eder, Ray Shapp, Flemming R. Ovesen, Torben Taustrup, and Bruce Card. This work was supported in part by NASA (grant NNX08AJ48G), the NSF (grant AST-0908357), and a Research Fellowship of the Humboldt Foundation. This research made use of the *SAO/NASA Astrophysics Data System*, the *Yahoo! Web Search* service, the Jet Propulsion Lab *Horizons* service, the *Python* programming language, and open-source software in the *numpy*, *scipy*, *matplotlib*, *MarkovPy*, *pYsearch*, and *WCSlib* projects. All code and data used in this project are available from the authors upon request.

REFERENCES

Ball, R. S., 1915, *A Popular Guide to the Heavens*, (Van Nostrand)

Buzzi, L., Muler, G., Kidger, M., Henriquez Santana, J. A., Naves, R., Campas, M., Kugel, F., & Rinner, C., 2007, IAU Circ., 8886, 1

Calabretta, M. R., & Greisen, E. W. 2002, A&A, 395, 1077

Gauthier, A., Christensen, L. L., Hurt, R. L. & Wyatt, R., 2008, in *Communicating Astronomy with the Public*, eds. L. L. Christensen, M. Zoulias, & I. Robson, 214

Giorgini, J.D., Yeomans, D.K., Chamberlin, A.B., Chodas, P.W., Jacobson, R.A., Keesey, M.S., Lieske, J.H., Ostro, S.J., Standish, E.M., & Wimberly, R.N., 1996, BAAS, 28(3), 1158

Goodman, J. & Weare, J., 2010, Comm. App. Math. and Comp. Sci., 5, 65

Gould, A., 2008, in *Manchester Microlensing Conference*, eds. E. Kerins, S. Mao, N. Rattenbury, & L. Wyrzykowski (SISSA), 38

Hedstrom, L., 2007, *pYsearch 3.0*: Python APIs for *Yahoo!* search services, `http://pYsearch.sourceforge.net/`

Hogg, D. W., Myers, A. D., & Bovy, J., 2010, ApJ, 725, 2166

Kepler, J., 1609, *Astronomia Nova*, trans. W. H. Donahue, 1992 (Cambridge University Press)

Korpela, E. J., *et al.*, 2009, in *Bioastronomy 2007: Molecules, Microbes and Extraterrestrial Life*, eds. K. J. Meech, J. V. Keane, M. J. Mumma, J. L. Siefert, & D. J. Werthimer (ASP) 420, 431

Lang, D., Hogg, D. W., Mierle, K., Blanton, M., & Roweis, S., 2010, AJ, 139, 1782

Lintott, C., *et al.*, 2011, MNRAS, 410, 166

This preprint was prepared with the AAS LaTeX macros v5.2.

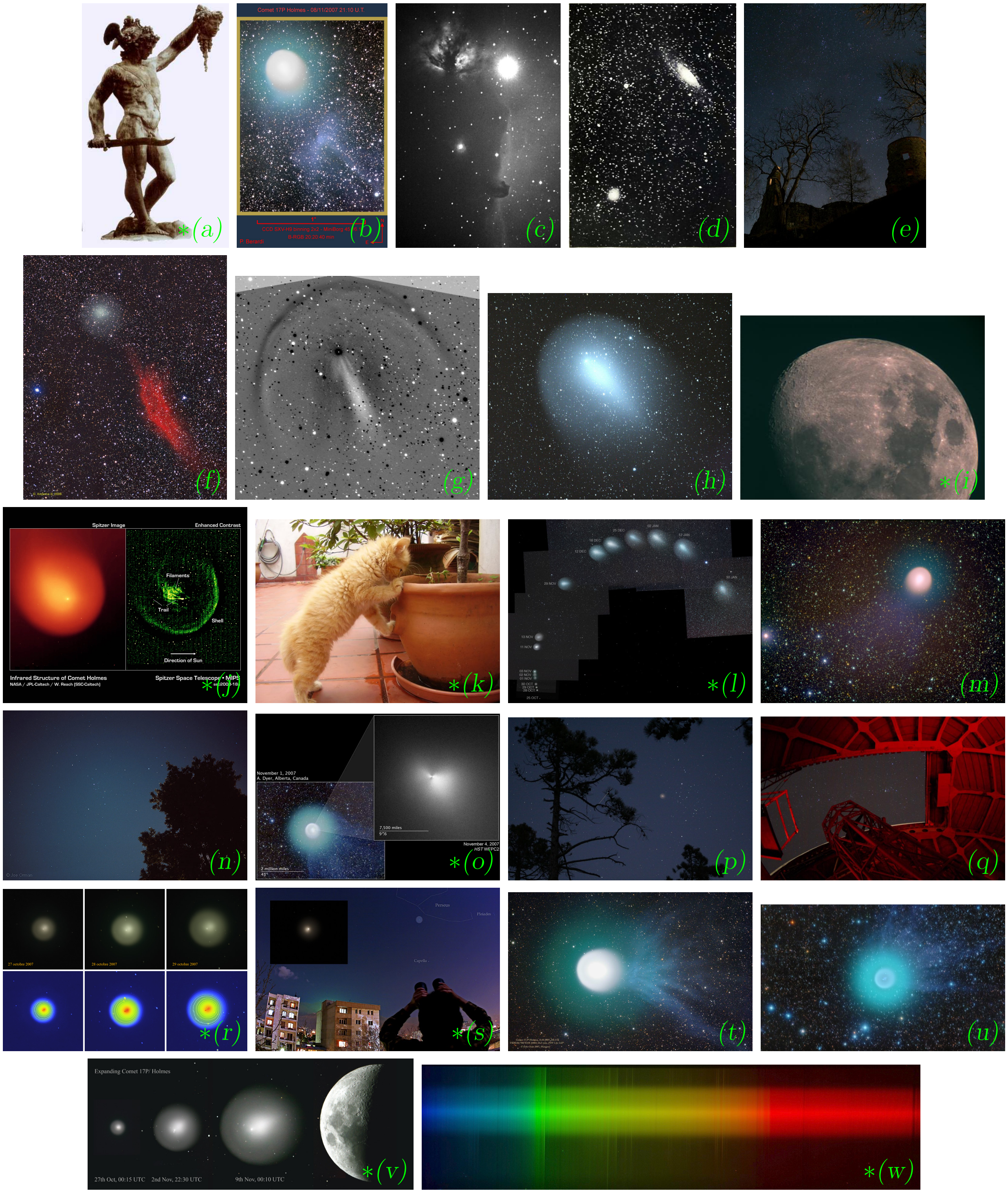

Fig. 1.— Example images. . . [caption next page]

Fig. 1.— [figure previous page] Example images from the image search. Here we have highlighted the diversity of the images; the majority are in fact high-quality, narrow-field images of the comet, such as *(b)*, *(h)*, *(m)*, and *(u)*. Images that did not calibrate successfully with *Astrometry.net* (and therefore were not used as data in this study) are marked with asterisks. Notice that images *(e)*, *(p)*, and *(q)* all were successfully calibrated and were used in the analysis. Image *(a)* shows a statue of Perseus, a constellation through which Comet Holmes travelled during its 2007 approach. Image *(f)* includes the California nebula (NGC 1499). We recommend a caption of "Iz commut in heer?" for image *(k)*. Credits: *(a)* copyright 2000–2005 Gods, Heroes, and Myth (`http://www.gods-heros-myth.com`); *(b)* Paolo Berardi; *(c)* Amateur Astronomers, Inc. Research Committee (`http://asterism.org`); *(d)* Edward Emerson Barnard (Ball 1915); *(e)* copyright P.-M. Hedén (`http://www.clearskies.se`); *(f)* copyright Dave Kodama (`http://astrocamera.net`); *(g)* TOC Observatory (`http://tocobs.org`); *(h)* copyright Fay Saunders; *(i)* Bruce Card, Aldrich Astronomical Society, Worcester MA; *(j)* NASA, JPL-Caltech, W. Reach (SSC-Caltech); *(k)* copyright Julián Cantarelli; *(l)* copyright 2007, 2008 John F. Pane (`http://holmes.johnpane.com`); *(m)* copyright Tyler Allred (`http://allred-astro.com`); *(n)* Joe Orman; *(o)* NASA, ESA, and H. Weaver (The Johns Hopkins University Applied Physics Laboratory), and A. Dyer, Alberta, Canada; *(p)* copyright Vincent Jacques (`http://vjac.free.fr/skyshows`); *(q)* Jimmy Westlake, Colorado Mountain College; *(r)* Stephane Zoll (`http://astrosurf.com/zoll`); *(s)* Babak Tafreshi / TWAN (`http://twanight.org`); *(t)* Ivan Eder (`http://eder.csillagaszat.hu`); *(u)* Vicent Peris (OAUV), José Luis Lamadrid (CEFCA); *(v)* copyright Thorsten Boeckel (`http://tboeckel.de`); *(w)* D. J. Barry, Department of Astronomy, Cornell University.

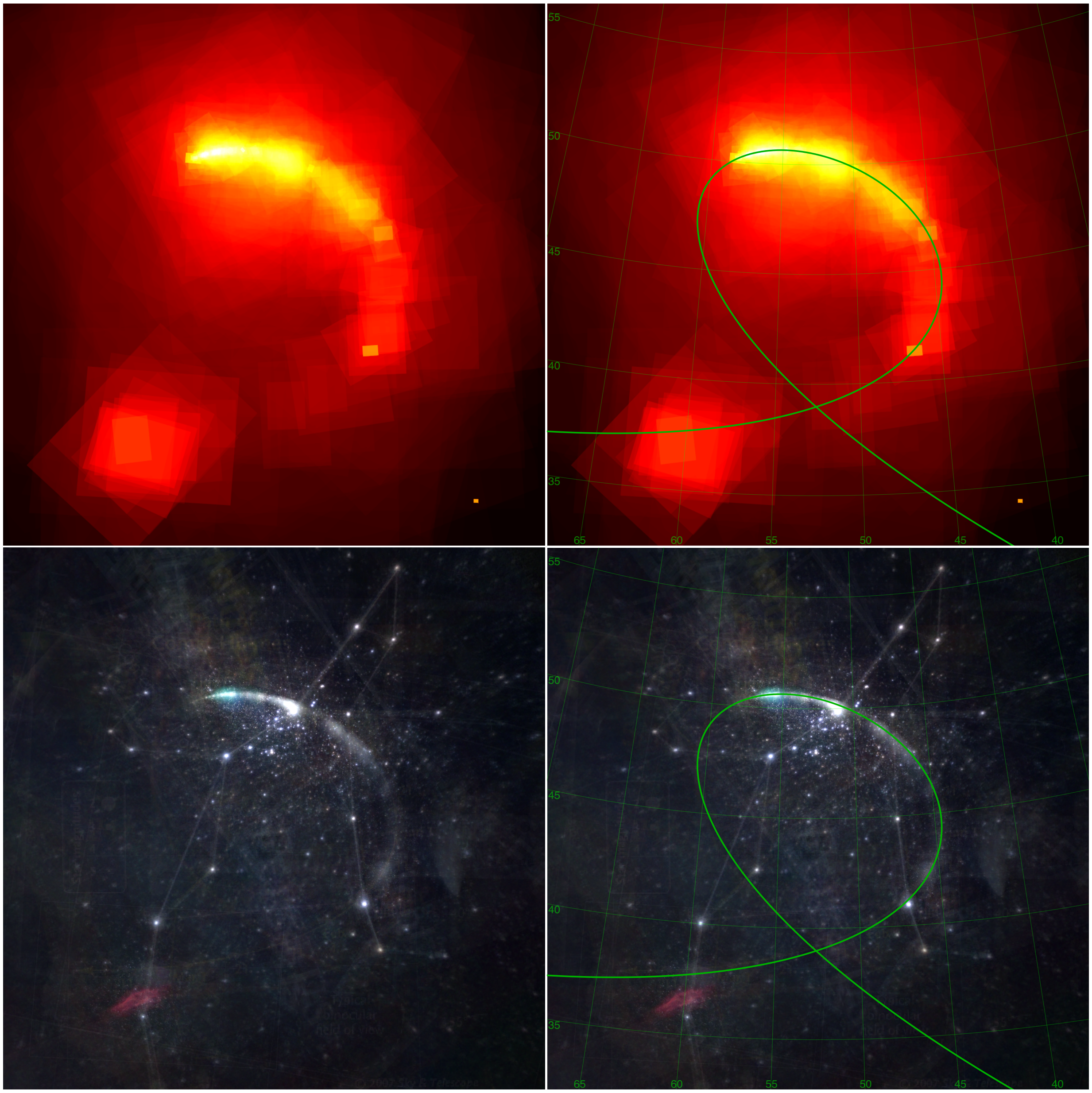

Fig. 2.— Images that were successfully calibrated by *Astrometry.net*, aligned in celestial coordinates. *Top:* Total pixel density of the images, with a log stretch. The most heavily imaged point on the sky is covered by 493 images and has a pixel density of over 4 million pixels per square degree. The right panel shows the same image as the left panel but with a coordinate grid and the trajectory of the JPL ephemeris (Giorgini *et al.* 1996) for Comet 17/P Holmes. *Bottom:* Co-added images. The co-added images show the fixed stars because the images have been aligned in celestial coordinates; they show some faint lines joining the stars because some of the images used in this study are diagrams of the constellations rather than just simple photographs. The California Nebula (NGC 1499) is visible in the bottom-left of the image because many photographers imaged the conjunction of the comet and nebula (see, for example, Figure 1*(f)*).

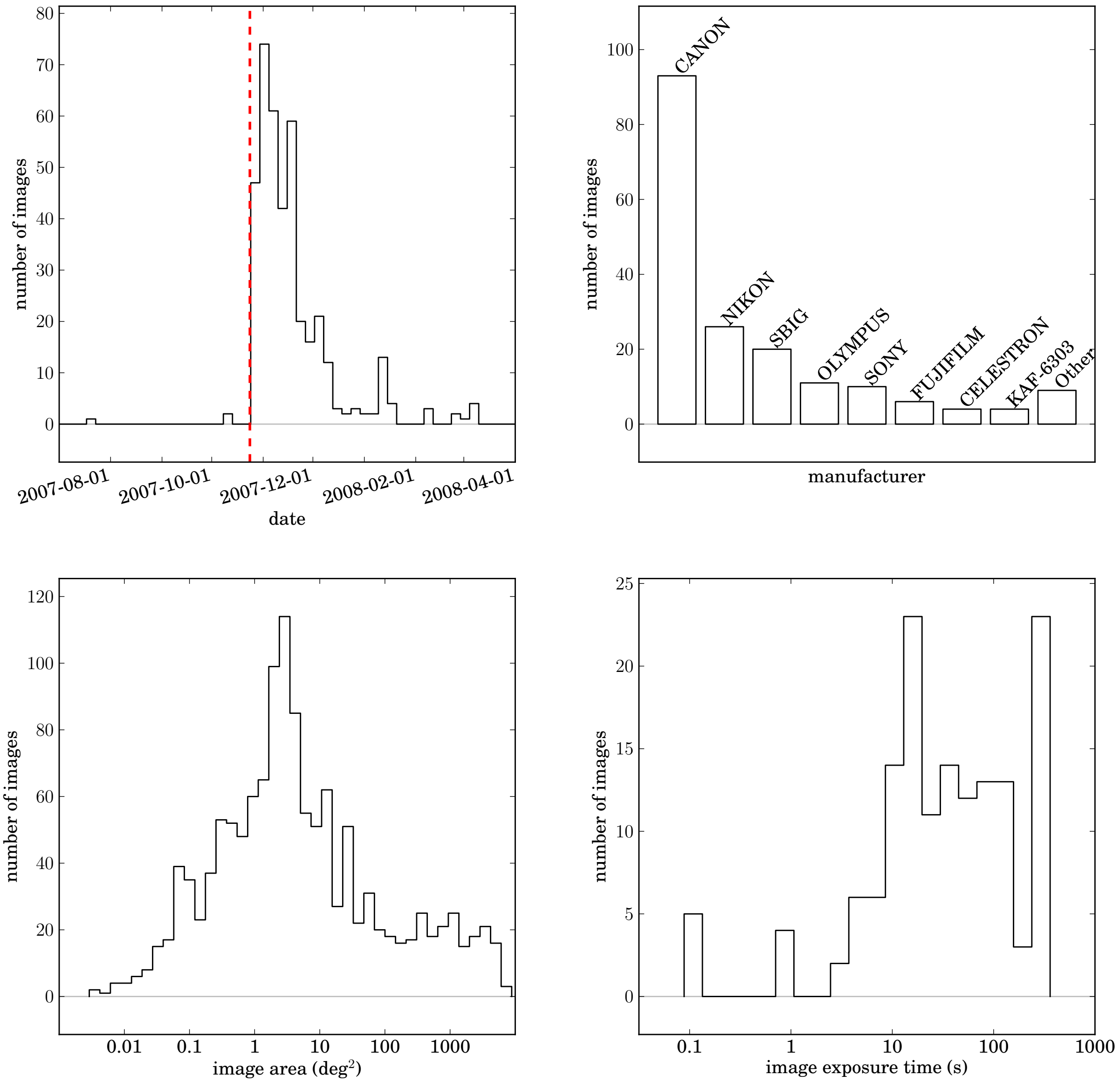

Fig. 3.— *Top-left:* The timestamps in the EXIF headers of the images in our data set. 422 of the 1299 images have timestamps. The dashed line marks 2007 Oct 24, the date of the comet's outburst and dramatic brightening (Buzzi *et al.* 2007). *Top-right:* The distribution of camera manufacturers listed in the EXIF headers. *Bottom-left:* The distribution of image angular sizes in our data set, according to *Astrometry.net*. *Bottom-right:* The distribution of exposure times in our data set, according to EXIF entries.

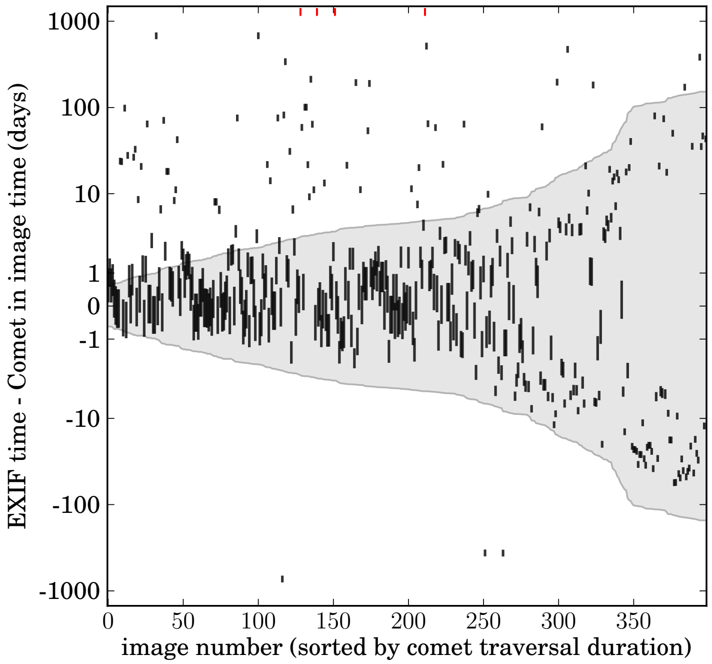

Fig. 4.— Evaluation of the accuracy of the timestamps in the image EXIF headers. For each image, we computed the range of times that the comet appeared inside the celestial-coordinate bounds of the image; that is the shaded gray region. The images are sorted so that this envelope is monotonic. The EXIF timestamp for each image is shown as a bar of height one day (since EXIF has no time zone specification, this is our intrinsic uncertainty), plus a minimum size to make all the markers visible. If the EXIF timestamps were correct and set to UT, the position of the bar within the gray region would indicate the position of the comet within the image. Most of the bars touch the gray region, indicating that the majority of the EXIF timestamps are consistent (that is, the comet would indeed appear within the image at the stamped time), and the inconsistent times are almost all *later*; perhaps these timestamps mark times at which the image was edited. The apparent bifurcation toward the right side of the plot is due to the arcsinh plot stretch.

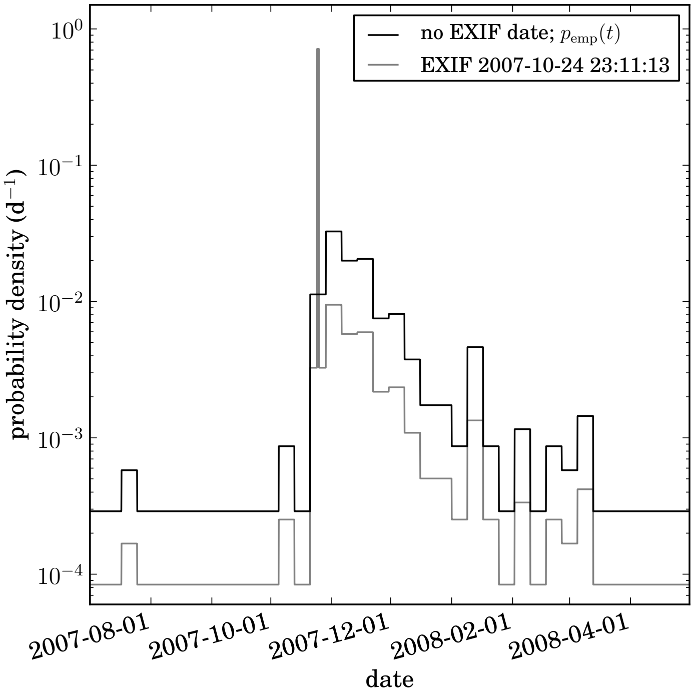

Fig. 5.— The time prior PDF $p_{\rm emp}(t)$ used for images with no EXIF date information (dark line) and the time prior PDF for an image with a particular EXIF date (lighter line). The latter is a mixture of $p_{\rm emp}(t)$ and a top-hat of width one day centered on the EXIF date and fractional weight $p_{\rm EXIF} = 0.71$. We use a width of one day because the EXIF standard does not permit the time zone to be specified. The function $p_{\rm emp}(t)$ is based on the distribution of reported EXIF dates shown in Figure 3; details in the text.

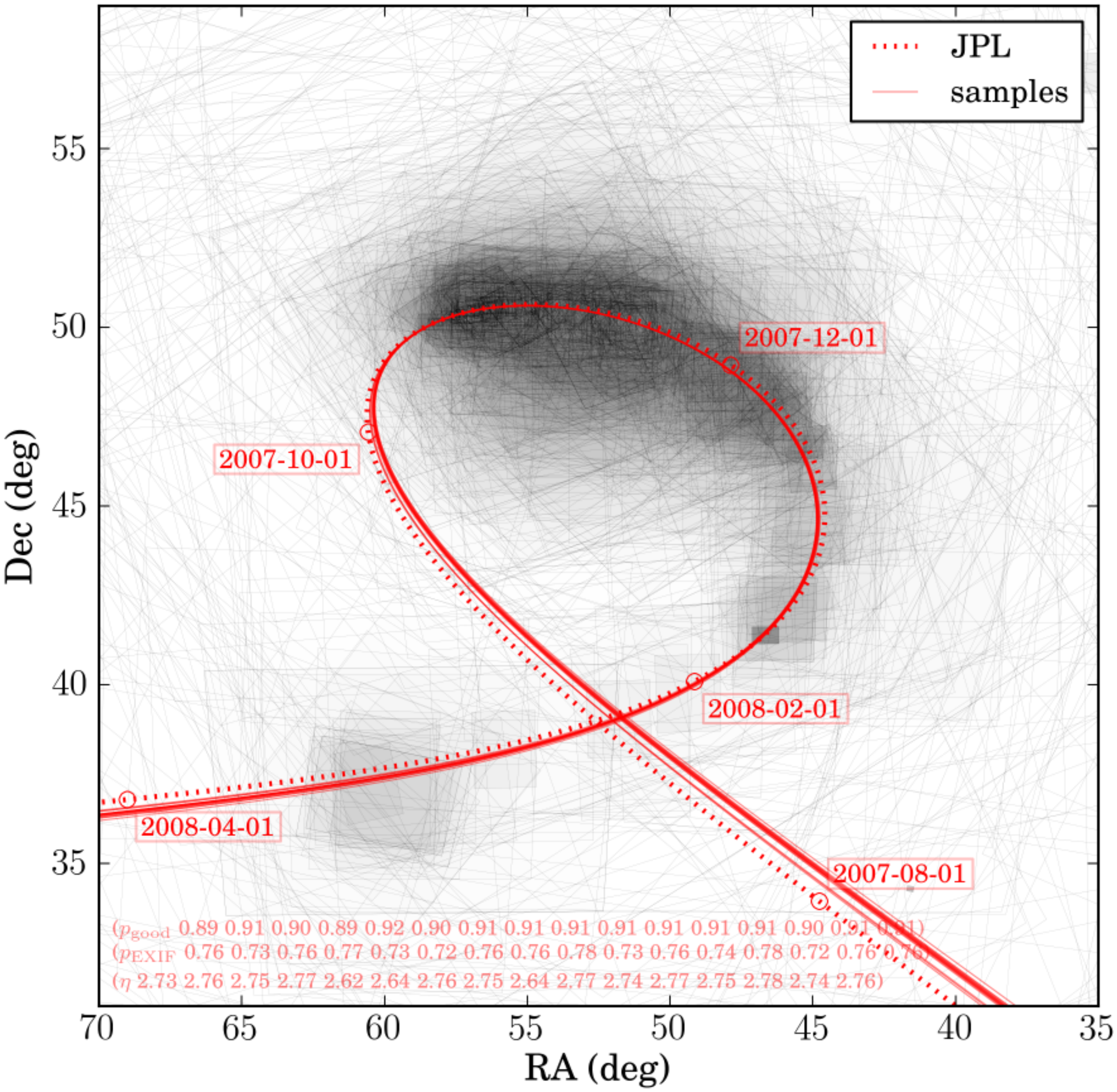

Fig. 6.— Image footprints with JPL ephemeris and inferred orbits superimposed. The solid lines show 16 samples from the posterior PDF for the parameters. The values for the hyperparameters $p_{\rm good}$, $p_{\rm EXIF}$, and η (see text for definitions) for the 16 samples are given in faint red at the bottom of the plot. In the background is shown the outlines of all the successfully calibrated image footprints.

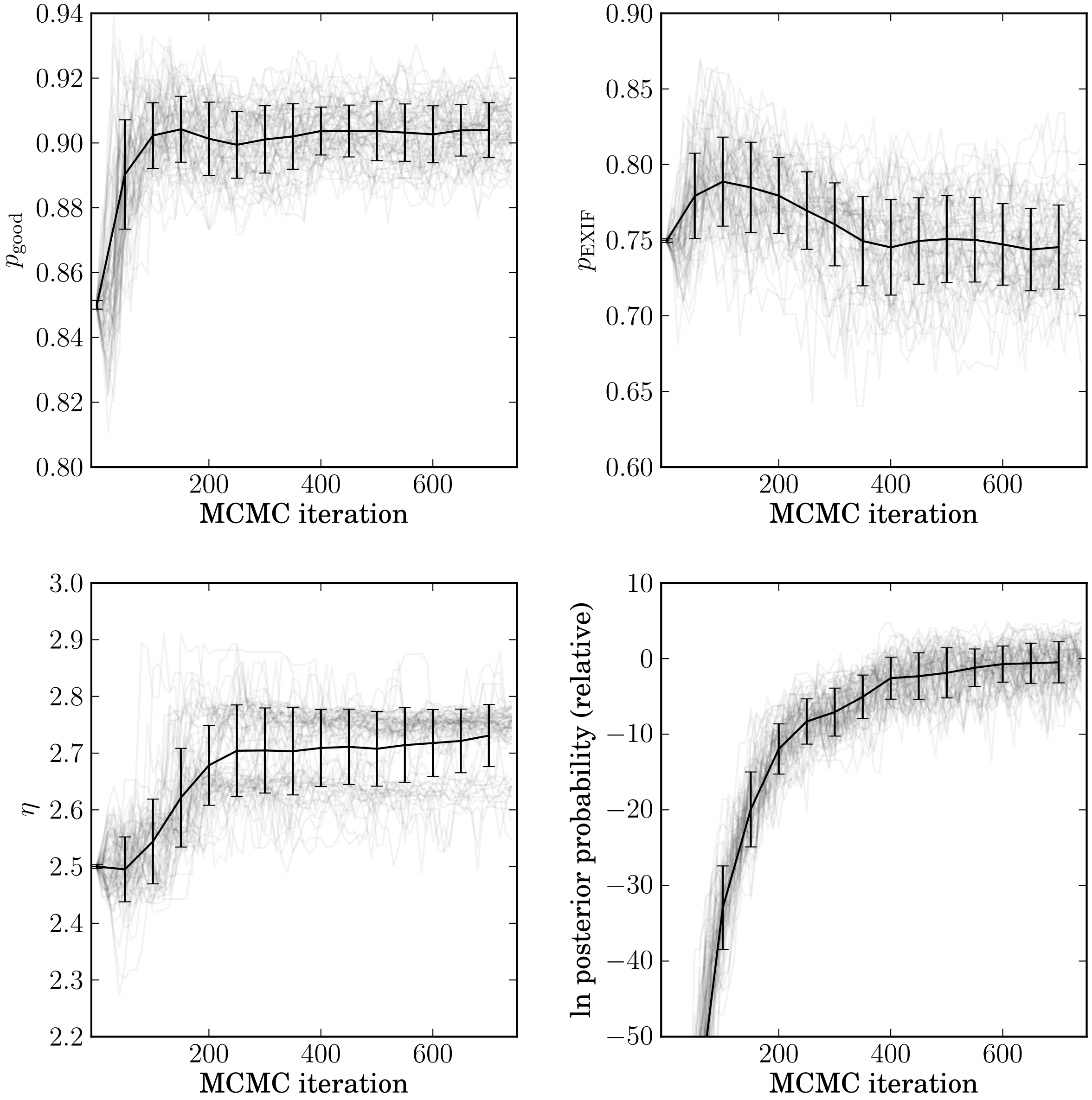

Fig. 7.— Hyperparameter values as the MCMC proceeded. *Top-left:* $p_{\rm good}$, the probability that an image is an image of the comet (that is, was generated by the foreground probability distribution $p_{\rm fg}$). *Top-right:* $p_{\rm EXIF}$, the probability that a timestamp in an image EXIF header is correct. *Bottom-left:* η, the (inverse of the) central fraction of the image area in which the comet appears. *Bottom-right:* the log posterior probabilities of the data given the parameter and hyperparameter values. The solid line and error bars show the mean and standard deviation of the 64 walkers in our ensemble; the individual values are shown with faint lines. After 500 iterations, the distributions have largely settled.

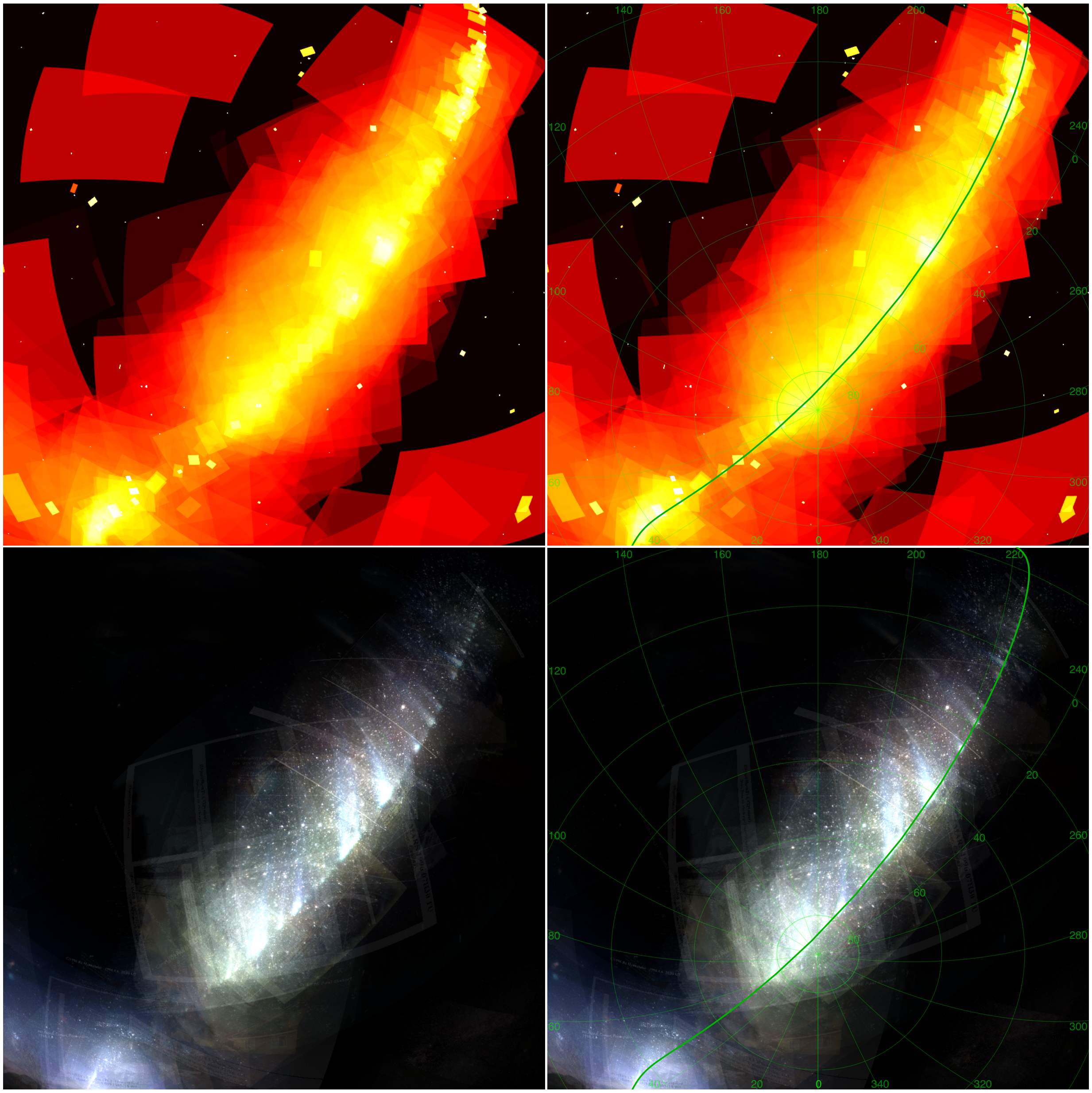

Fig. 8.— Same as Figure 2 but now for images found by a similar Web search for “Comet Hyakutake”. The search, performed on 2010 Oct 2, produced 1481 JPEG images, of which 1019 were recognized by *Astrometry.net* as images of the night sky. *Top:* Pixel density map, with a log stretch. The most heavily imaged areas appear in 152 images. The projection is zenithal equidistant (FITS WCS code “ARC”, Calabretta & Greisen 2002). *Bottom:* Co-added images. The spectacular tail of the comet is clearly visible, as are many text labels, annotations, and image borders. The right panels show the same images as the left panels but with a coordinate grid and the trajectory of the JPL ephemeris.